What lies beneath?

Janzen, Thijs

Chapter 6

The role of habitat dynamics in driving diversification

Thijs Janzen, Rampal S. Etienne

ABSTRACT

It is commonly accepted that geographical isolation can play an important role in speciation. Geographic isolation is often assumed to slowly increase over time, for instance through the formation of rivers, mountains or the movement of tectonic plates. Cyclic changes in connectivity between areas might occur however when water levels fluctuate in a large lake, or when changes in sea level changes the connectivity between islands. These habitat dynamics may act as a driver of allopatric speciation and propel local diversity. Here we present a basic model of this interaction between changes in the environment and speciation. We model fluctuations in water level and compare results of our model with a published phylogeny of cichlid fish from Lake Tanganyika where such cyclic changes have occurred. Simulation data confirm that our model is able to recover water level changes from phylogenies. When confronting our model to the phylogeny of cichlid fish from Lake Tanganyika, we do not find evidence for water level changes and associated allopatric speciation. This suggests that large-scale water level fluctuations have had little impact on the current diversity of cichlids in Lake Tanganyika. However, we argue that the Yule tree model prior used to reconstruct the phylogeny may have biased our results, and therefore advocate the incorporation of more complex tree model priors that take into account habitat dynamics.

Introduction

Without speciation, ultimately all diversity will be lost. Speciation is the process in which from one population, two sub-populations form which are reproductively isolated (Coyne & Orr 2004). There are two main trajectories leading to reproductive isolation. Firstly, reproductive isolation can be driven by sexual selection or preferential mate choice. The formation of reproductive isolation by sexual selection takes place in full sympatry, and does not require the two sub-populations to be spatially isolated. Secondly, reproductive isolation can be driven by natural selection: speciation driven by environmental factors that cause a divergence in traits between the two populations. Although spatial isolation is not fully necessary, the classical view is that of two populations that have become isolated from each other due to a geographical change in the environment such as the formation of a mountain ridge. Traits in the two populations diverge from each other, either because of genetic drift, or because of local adaptation to different conditions.

Evidence for the classic view of slowly changing environments causing spatial isolation and ultimately speciation is ample (Coyne & Orr 2004). Environmental changes such as the formation of mountain ridges, the formation of rivers and the movement of tectonic plates have been shown to have influenced speciation processes and thereby biodiversity. Nevertheless, the long timescales upon which these processes operate leave room for speculation whether the high levels of biodiversity we observe today was solely the result of interactions between these slow environmental changes and opportunities for speciation. Recently, cyclic changes in the environment, called "landscape dynamics" (Aguilée *et al.* 2009, 2011), have been recognized as a possible alternative driver of diversity, changing our view of the speed and dynamics of speciation under natural selection.

Cyclic changes in the environment can cause populations to continuously switch between allopatric and sympatric stages, providing a continuously renewed potential for speciation. Connections between populations may change due to glaciations and postglacial secondary contacts (Barnosky 2005 and references therein). Alternatively islands may become fragmented or undergo fusion as a result of sea level changes (Glor *et al.* 2004; Thorpe *et al.* 2008). In a similar fashion, fluctuations in water level can cause fragmentation and fusion of lakes, as has happened frequently in the African Rift Lakes.

The speed at which the habitat dynamics operate determines the amount of diversity in the system. If the biogeography changes too fast, there is not enough time for the isolated populations to be influenced by natural selection and to adapt to the different habitats they experience in isolation, or to diverge due to genetic drift. After secondary contact, the lack of divergence between the populations will reduce opportunities for reproductive isolation, and speciation is unlikely. If on the other hand habitat dynamics are too slow, the total number of species will be relatively low, as few opportunities for population subdivision have presented over time. Timing of the habitat dynamics might therefore play a crucial role in how radiations proceed through time.

Aguilée and coworkers developed a model in which populations at different locations diverge from each other depending on the local habitat, and at the same time allow for sympatric speciation by implementing a standard assortative mating system which allows for a single branching point in trait values (Aguilée *et al.* 2013). As the different locations become separated or connected over time, new species form. The authors conclude that stable numbers of diversity are best obtained by having a fragmented habitat with recurrent merged states and rapid fluctuations. Although their model is based on the water level fluctuations in the African Rift Lakes, the authors do not compare their results with diversity levels in these lakes.

Confronting models of diversification through habitat dynamics with known phylogenies may allow us to detect signatures of historical habitat dynamics in these phylogenies, and to assess their importance. Using a spatially explicit model of landscape fragmentation, Pigot and coauthors (Pigot *et al.* 2010), compared phylogenies generated heir model with known bird phylogenies. In their spatially explicit model, species occupy a geographic range. Consecutive splitting of these geographic ranges drives speciation in the model, and including the geographical context of speciation can explain a large part of the features exhibited by the reconstructed avian trees.

The models of Aguilee *et al.* (2013) and Pigot *et al.* (2010) show that the geographical background of speciation is an important factor to take into account. Currently, no model has compared the influence of cyclic habitat dynamics on phylogenies. Although the model by Pigot *et al.* 2010 compares characteristics of bird phylogenies with phylogenies generated by the model, their model mainly looks at fragmentation of habitat ranges, and does not incorporate the fusion of habitat ranges. The model of Aguilée *et al.* 2013 in contrast does incorporate cyclic habitat dynamics, but their analysis remains conceptual, and they do not apply their model to empirical data. Here we present a model that explicitly models diversification as a result of cycles of fragmentation and fusion of the environment, and we infer diversification rates from empirical data.

The African Rift Lakes provide a good starting point in studying the interplay between habitat dynamics and speciation. The African Rift Lakes are known to have been subject to frequent water level changes (Cohen *et al.* 1997b; Alin & Cohen 2003), which might have influenced fish diversity in the lakes. An estimated number of 2000 cichlid fish species have evolved over the past 10 million years (Turner *et al.* 2001), which is considered one of the most spectacular adaptive radiations (Seehausen 2006). The most prominent water level changes have taken place in Lake Tanganyika, where the water level dropped over 600 meters, splitting the lake into multiple smaller lakes (Lezzar *et al.* 1996; Cohen *et al.* 1997a, 2007). Being the oldest lake of the three large rift lakes (Cohen *et al.* 1993), it contains the highest behavioural diversity (Konings 2007) and it is the only lake with a highly resolved phylogeny for cichlid fish. Microsatellite analysis points out that in several genera, we find populations that are aligned with the locations of the smaller lakes at low water level, suggesting the influence of water level changes on population segregation and speciation.

Here, we ask how cyclic changes in the environment influence both the generation and the maintenance of biodiversity. Secondly we aim to extract information about past habitat dynamics from known phylogenies, by comparing known phylogenies with the outcomes of a model that describes the influence of habitat dynamics on diversity. This model is an extension of the standard constant-rates birth-death model. Phylogenies generated by the model are compared with the phylogeny of the *Lamprologini*, a tribe of cichlid fish from Lake Tanganyika. We both directly infer past water level changes, and make use of literature findings of past water level changes in order to assess the importance of these habitat dynamics in shaping the current biodiversity of cichlids in Lake Tanganyika.

Methods

Let us consider a lake that subdivides into two pockets when the water level drops. Water level changes are stochastic events and we only consider water level changes that result in a water drop that is significant enough to split the lake into two pockets. When the water level drops, we assume that all species distribute themselves equally over the two pockets; similarly, when the water level rises, all species previously contained in one pocket distribute themselves over the entire lake. Allopatric speciation can only occur when the water level is low. After a waiting time, which is a stochastic variable in our model, one of the two incipient species can speciate into a new species. If this allopatric speciation does not occur before the water level rises again, i.e. reflecting that there has not been enough genetic divergence, the two incipient species in the two pockets are merged back into one species again. Sympatric speciation can always occur, either at high water level in the lake, or in both pockets when the water level is low. The sympatric speciation rate is not modified when the water level drops, and hence at low water level, the potential for sympatric speciation is twice as high. Extinction is considered to be a background process that occurs locally, i.e. in a pocket. . If the water level is high, this causes extinction of a species, if the water level is low, this causes local extinction in one of the pockets.

We inplemented our model using the Gillespie algorithm, where the time steps are chosen depending on the rate of possible events. In the current model there are four different possible events:

1) A water level change event, with rate ω
2) Sympatric speciation event, with rate σ N
3) Allopatric speciation event, with rate αP, if the water level is low, otherwise 0
4) Extinction event, with rate μN

where N is the total number of (incipient) species, and P is the number of species that is present in both pockets.

When the population is distributed over two pockets, essentially all species are copied, such that now we have twice as many instances of species (this affects the

sympatric speciation rate, extinction rate and allopatric speciation rate, because *N* is doubled). Thus, immediately after a water level drop, the number of incipient species *N* is equal to 2*S*, where *S* is the number of species. When the populations are merged, all incipient species that belong to the same species are merged to a single species, such that after a water level rise *N* is equal to *S*, where *S* is the number of unique species. During a sympatric speciation event, a single species splits into two new species, and the original (incipient) species is consumed in the process. This implies that if the water level is low, the incipient species that is sister/counterpart to the ancestral species might survive in the other pocket.

During an extinction event, one (incipient) species is removed from the simulation. If the water level is low, this need not lead to the extinction of a species, as the sister incipient species might remain in the other pocket.

Empirical data

We fitted our model to the phylogenetic tree of the tribe of *Lamprologini* (Sturmbauer *et al.* 2010). The tribe of *Lamprologini* is the most diverse tribe within Lake Tanganyika, and contains about 80 species of cichlids (Koblmüller *et al.* 2007; Day *et al.* 2007; Sturmbauer *et al.* 2010). The *Lamprologini* are an endemic tribe of Lake Tanganyika, and all species are substrate brooders with shared paternal and maternal care. In contrast to the mouthbrooding species from the *Haplochromini*, which can be found in all three lakes, the *Lamprologini* show little sexual dimorphism and dichromatism, which are well known indicators for sexual selection (Kraaijeveld *et al.* 2011). We therefore expect that if changes in water level have driven allopatric speciation events, the tribe of *Lamprologini* is the tribe where we are most likely to pick up any signals from these past events, as diversity in this tribe seems to be less driven by sexual selection.

We used the concensus *Lamprologini* tree published by Sturmbauer *et al.* (2010). This tree is based on the complete ND2 sequences of 91 species, which represents 79 lamprologine species and 12 outgroup taxa (Sturmbauer *et al.* 2010). Tree topology was the consensus tree based on topologies obtained using neighbor joining (NJ), Maximum Likelihood (ML) and Bayesian Inference (BI). Divergence times were estimated in BEAST (Drummond *et al.* 2012) using a SRD06 two-partition codon-specific rates model, using a relaxed molecular clock with log-normally autocorrelated rates among branches (Sturmbauer *et al.* 2010). All priors were left at default settings, except for the tree prior which was set to be a Yule prior. The tree was calibrated assuming that the diversification of the tribe of *Lamprologini* coincided with the formation of tropical clearwater habitat with deep-water conditions, 5-6 MYA (Salzburger *et al.* 2002b), the time window for the Congo River *Lamprologus* to leave Lake Tanganyika via the Lukuga river (Lezzar *et al.* 1996) and using the age of the Lake Malawi species flock (Delvaux 1995; Sturmbauer *et al.* 2001). In order to focus solely on diversification patterns in the tribe *Lamprologini* we pruned the tree from outgroups using the package APE (Paradis *et al.* 2004) in R. We removed any riverine species (*N. devosi*,

L. congoensis & *L. teugelsi*). Furthermore we selected only one sample for those species with multiple entries, namely for *T. temporalis*, *J. ornatus*, *P. toae*, *N. similis*, *N. leloupi*, *N. savoryi* and *T. dhonti*.

Maximum Likelihood

Without water level changes, our model reduces to a standard birth-death model (Nee *et al.* 1994). As a reference therefore, we estimated parameters of the standard birth-death model using Maximum Likelihood. The likelihood of the birth-death model was calculated using the function "globalBiDe.likelihood" from the package TESS (Höhna 2013a). Maximum Likelihood optimization was performed using the package subplex (King 2014) .

Fitting the model to empirical data

To fit the model to empirical data we made use of Approximate Bayesian Computation, in combination with a Sequential Monte Carlo scheme (ABC-SMC). As summary statistics for the ABC analysis we chose the normalized LTT statistic (Janzen *et al.* 2014), tree size, Phylogenetic Diversity (Schweiger *et al.* 2008) and the γ statistic (Pybus & Harvey 2000).

On all four parameters (ω, σ, α, ε) we chose uniform priors U(-7,3), on a ^{10}log scale, such that the eventual prior distribution spans (10^{-7},10^{3}). The jumping distance was 0.05 (on the ^{10}log transformed parameter), and we updated one parameter each time (e.g. jumps were only made in one dimension, to avoid extremely low acceptance rates).

The ABC-SMC scheme proceeds as follows:

1. Initialize $\varepsilon_1 \ldots \varepsilon_T$
   Set the population indicator $t = 0$
2. Set the particle indicator $i = 1$.
3. If $t = 0$, sample θ^{**} from the prior π
   Otherwise sample θ^{*} from the previous population θ^{i}_{t-1} with weights w_{t-1} and perturb the particle to obtain θ^{**} such that $\theta^{**} = \theta^{*} + N(0,0.05)$
4. If $\pi(\theta^{**}) = 0$, return to 3.
5. Simulate a tree $T^{*} \sim \theta^{**}$
6. Calculate summary statistic S for tree T*
7. If $|SS(T^{*}) - SS(T_0)| > \varepsilon_t$ return to 3.
8. Set $\theta^{i}_{t} = \theta^{**}$ and calculate the weight for particle θ^{i}_{t}:
   If t = 0; $W^{i}_{t} = 1$

   If t > 0; $$W^{i}_{t} = \frac{\pi(\theta^{i}_{t})}{\sum_{j=1}^{N} w^{j}_{t-1} N(\theta^{j}_{t}, \theta^{i}_{t})}$$
9. If $i < N$, set $i = i + 1$, go to 3.
10. Normalize the weights
11. If t < T, set t = t + 1, go to 2.

The number of particles was set to 10,000. Threshold values for the different summary statistics were chosen to decrease exponentially with every iteration, such that $\varepsilon_t[t] = \varepsilon_0 * \exp(-0.25 * t)$, where t is the indicator for the iteration. Initial ε values were 0.2, 50, 1 & 1 for the normalized LTT statistic, tree size, the Phylogenetic Diversity and the γ statistic respectively. As distance metric for the latter three summary statistics we used the absolute difference between the summary statistic of the simulated tree and the summary statistic of the empirical tree. Summary statistic values for the empirical tree where 76 (number of taxa), 0.95 (Phylogenetic Diversity) and -2.07 (γ statistic). To compare nLTT curves, we used the absolute distance, as in equation (1) in Janzen *et al.* (2014).

Water level change implementations

Lake Tanganyika experienced low water level stands 35–40k years ago, 169–193kya, 262–295kya, 363–393kya and 550–1100kya (Lezzar *et al.* 1996; Cohen *et al.* 1997a). Consequently, high water levels occurred between 0–35kya, 40–169kya, 193–262kya, 295–363kya and 393–550kya. Unfortunately the geological record does not reveal if any low water level stands occurred beyond 1 million years ago. This leaves a considerable gap in our knowledge about past water level stands, considering that the phylogeny of *Lamprologini* spans 5.28 My. We therefore used the literature values to extrapolate changes in water level beyond 1 million years ago in two different ways.

Firstly the waiting time until the next water level change was randomly selected from either the known periods of low water level stand if the water level was low (e.g. either 5k, 24k, 30k, 33k and 550k years) or randomly picked from the known periods of high water level stand if the water level was high (e.g. either 35k, 68k, 69k, 129k and 159k years). We did not use the literature values directly, as this would only provide information for the past 1 million years, and would not solve the problem of extrapolating to the full 5.28 My. In order to have a consistent distribution across all 5.28 My we chose to randomly pick a period from the literature values across all 5.28 My.

Secondly, we used the *average* number of water level changes over the past 1 million years. In the past 500,000 years, Lake Tanganyika experienced 4 low water level stands and thus experienced 8 water level changes, which translates to an average rate of 16 water level changes per million years. However, we know from the literature that Lake Tanganyika experienced only one more low water level stand in the previous 500,000 years, thus experiencing 5 low water level stands in the past 1 million years (and 10 water level changes). We used these two rates (10 and 16) as fixed rates in our model.

Alternatively, we did not use the literature values, but left the rate of water level change as a free parameter, to be inferred using the ABC-SMC approach. Because this approach has more degrees of freedom, we used 100,000 particles instead of the 10,000 particles used in all other approaches.

Model validation

Reliability of the model depends heavily on whether the model can accurately generate phylogenies, and secondly, whether parameter inference is accurate. When we remove water level effects from the model, the model reduces to the constant-rates birth-death model (Nee *et al.*, 1994). Thus, if our model can reliably generate and infer data, any inferences made on phylogenies generated without water level changes should match inferences made on these phylogenies using the constant-rates birth-death model. We therefore generated phylogenies using our model, and estimated parameters using the constant-rates birth-death model within a Maximum Likelihood framework and using our model within the ABC-SMC framework. We generated phylogenies using an extinction rate of -1, -2 or -3 (^{10}log), and a sympatric speciation rate of 0, -0.25 or -0.5 (^{10}log). For every combination we generated 5 phylogenies and for every phylogeny we estimated the sympatric speciation and extinction rate using MLE and using the ABC-SMC method described above, using 1,000 particles. We found that estimates using our model within an ABC framework were highly similar to estimates using the constant-rates birth-death model, thus confirming that our model accurately infers parameters in the special case where no water level changes are present.

Secondly we simulated trees for different combinations of sympatric and allopatric speciation with water level change. We either simulated phylogenies using sympatric speciation rates of 0, -0.25 or -0.5 (^{10}log) and water level rates of -1, 1 and 2 (^{10}log) or used allopatric speciation rates of 0, -0.25 or -0.5 (^{10}log) and water level rates of -1, 1 and 2 (^{10}log). Other parameters were set at 0. Using these phylogenies we tested whether our inference method can accurately infer these parameter values. We did not test combinations of all four parameters (extinction, sympatric speciation, allopatric speciation and water level rate) because these tests are highly computationally demanding.

To assess the uncertainty in the parameter estimates obtained for the empirical data we performed an *a posteriori* model validation test, where we sampled 1,000,000 parameter combinations from the posterior distribution for the four parameters (extinction, sympatric speciation, allopatric speciation and water level change). Using these 1,000,000 parameter combinations, we simulated the model and compared summary statistics of the simulated phylogenetic trees with the summary statistics found for the phylogeny of *Lamprologini*. If the empirical values fall outside the distribution of summary statistics, we can reject our model as a plausible explanation of *Lamprologini* diversification in Lake Tanganyika.

Results

Testing the model on simulation data

To test whether our inference method can detect water level changes at all, we first tested performance of our inference method on a number of simulated datasets.

Comparing to the constant-rates birth-death model

When water level rates are set to zero, allopatric speciation can no long occur, and the model reduces to a standard birth-death model, with a constant probability of speciation (sympatric speciation) and a constant probability of extinction. We find that the sympatric speciation rate is always estimated correctly, and inference of sympatric speciation rates is equal to obtained estimates using maximum likelihood (Figure 6.1 and Table 6.1). Extinction rates are less well estimated, but generally, estimates using

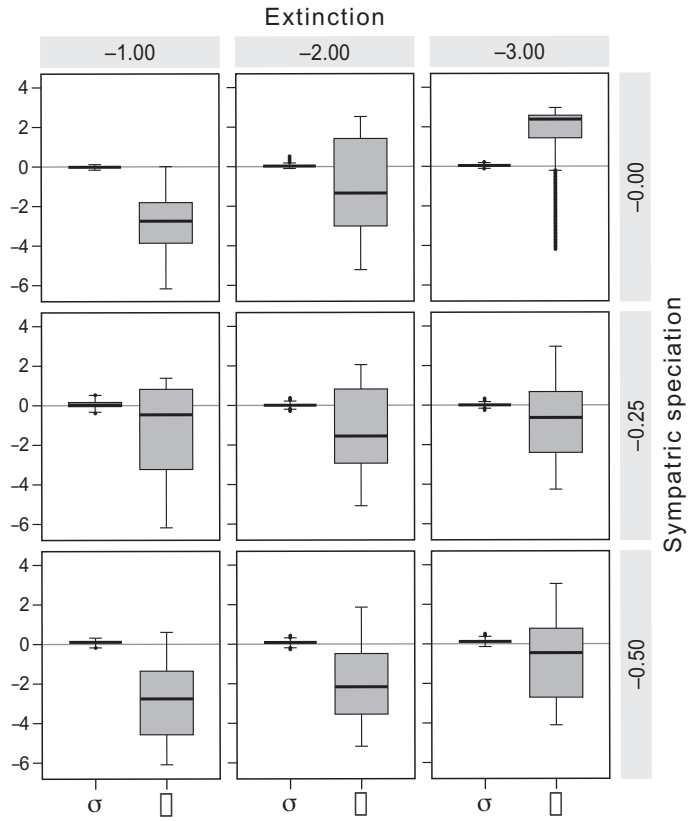

Figure 6.1. Boxplots of residuals of estimated parameter values for 9 different parameter combinations of ^{10}log(sympatric speciation rate) and ^{10}log(extinction rate). Estimates for the sympatric speciation rate (σ) tend to be good, whilst the extinction rate (μ) is generally underestimated. Number of replicates per parameter combination is 5. The grey line indicates residual = 0, e.g. the value with which the trees were generated.

ABC are, as expected, similar to values obtained using Maximum Likelihood on the constant-rates birth-death model (Table 6.1), for which it is known that extinction estimates can be inaccurate (Paradis 2004; Rabosky 2010; Aldous *et al.* 2011; Stadler 2013).

Table 6.1. Estimates for trees generated using 9 different parameter combinations. Both estimates using ABC-SMC, and estimates of the constant-rates birth-death model obtained using maximum likelihood are shown. Shown are means and standard deviations (between brackets) over 5 replicates.

		ABC estimate		ML estimate	
^{10}log (σ)	^{10}log (μ)	^{10}log(σ)	^{10}log (μ)	^{10}log (σ)	^{10}log (μ)
0	-1	-0.03 (0.04)	-4.03 (1.58)	-0.03 (0.04)	-4.44 (2.57)
0	-2	0.08 (0.14)	-2.97 (2.50)	0.08 (0.16)	-3.33 (2.97)
0	-3	0.05 (0.04)	-1.67 (2.08)	0.05 (0.04)	-0.76 (0.70)
-0.25	-1	-0.21 (0.18)	-2.35 (2.27)	-0.19 (0.22)	-0.62 (0.62)
-0.25	-2	-0.25 (0.10)	-3.29 (2.04)	-0.20 (0.08)	-3.02 (3.58)
-0.25	-3	-0.25 (0.07)	-3.73 (2.00)	-0.21 (0.05)	-2.20 (3.00)
-0.5	-1	-0.43 (0.09)	-3.83 (1.90)	-0.19 (0.42)	-1.03 (2.47)
-0.5	-2	-0.44 (0.09)	-3.99 (1.74)	-0.53 (0.16)	-5.10 (2.36)
-0.5	-3	-0.40 (0.10)	-3.61 (2.04)	-0.39 (0.18)	-4.59 (4.09)

Inferring water level rates from simulated data

First we assessed performance on inferring sympatric speciation and water level change, whilst keeping extinction and allopatric speciation zero. Estimates of sympatric speciation and water level change match the parameters used to generate data well (Figure 6.2 and Table 6.2). Especially for high levels of sympatric speciation estimates of both sympatric speciation and water level change are very accurate. When sympatric speciation rates decline, average tree sizes decline as well (because total time is kept constant at 5.28 million years, equal to the time of the *Lamprologini* tree). As tree size decreases, the total amount of information contained in the tree decreases, which is reflected in large variation of the parameter estimates and an underestimation of the number of water level changes.

Secondly we estimated allopatric speciation and water level change, whilst keeping sympatric speciation and extinction zero. Estimating allopatric speciation and water level change tends to be highly accurate as well, provided that water level changes are frequent enough to generate potential for allopatric speciation. Trees generated with low rates of water level change are very small on average, and consequently contain little information. Parameter inference for these trees tends to be inaccurate, as is reflected by the mean estimates (Table 6.3, $^{10}\log(\omega) = -1$), and the difference with the target parameter in Figure 6.3. When the water level change rate increases, tree sizes increase and accuracy of the estimates increases as well, even when allopatric speciation is relatively low (Table 6.3).

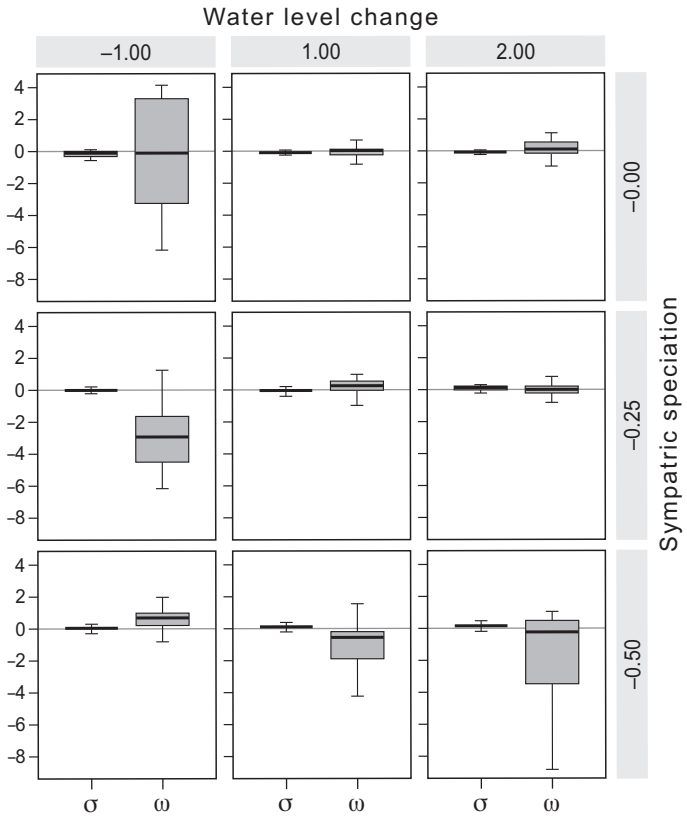

Figure 6.2. Boxplots of residuals of estimated parameter values for 9 different parameter combinations of ^{10}log(sympatric speciation rate (σ)) and ^{10}log(water level rate (ω)). Number of replicates per parameter combination is 5.

Table 6.2. Estimates for trees generated using 9 different parameter combinations of the sympatric speciation rate (σ) and the water level change rate (ω). The last column shows the tree size of the tree generated with the used parameter combination. Shown are means and standard deviation (between brackets) over 5 replicates.

^{10}log (σ)	^{10}log (ω)	^{10}log (σ)	^{10}log (ω)	Tree Size
0	-1	-0.11 (0.15)	-1.26 (3.22)	242.4 (140.67)
0	1	-0.07 (0.05)	0.98 (0.30)	183.6 (111.22)
0	2	-0.10 (0.05)	2.19 (0.41)	106.4 (36.78)
-0.25	-1	-0.29 (0.09)	-3.99 (1.79)	16 (10.10)
-0.25	1	-0.31 (0.10)	1.15 (0.40)	127.2 (75.09)
-0.25	2	-0.21 (0.11)	1.85 (0.66)	116 (86.21)
-0.5	-1	-0.47 (0.11)	-0.95 (1.86)	18 (13.56)
-0.5	1	-0.44 (0.12)	-0.66 (2.26)	20.8 (14.04)
-0.5	2	-0.38 (0.13)	0.55 (2.71)	27 (19.52)

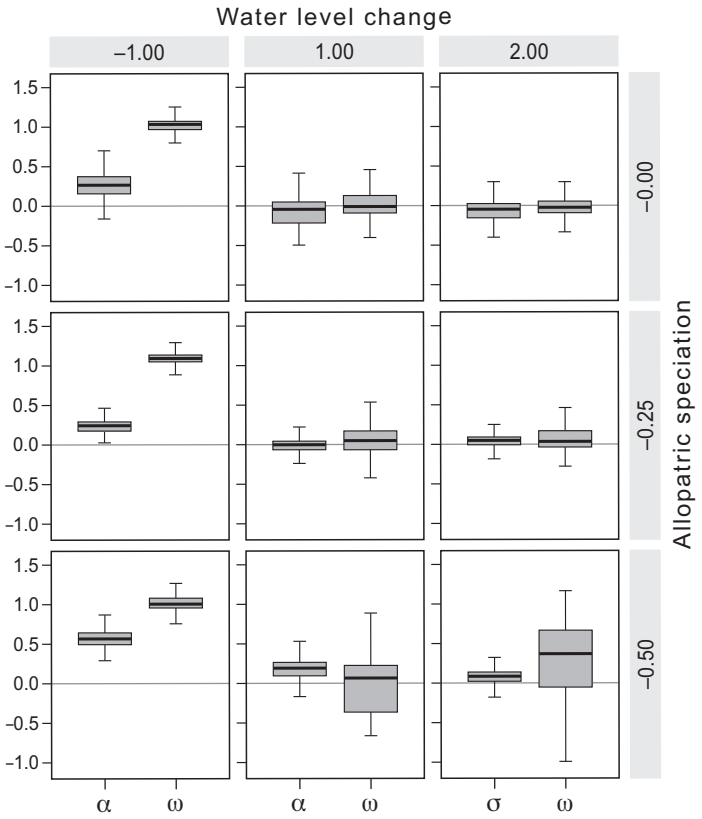

Figure 6.3. Boxplots of residuals of estimated parameter values for 9 different parameter combinations of ^{10}log(allopatric speciation rate (α)) and ^{10}log(water level rate (ω)). Number of replicates per parameter combination is 5.

Table 6.3. Estimates for trees generated using 9 different parameter combinations of the allopatric speciation rate (α) and the water level change rate (ω). The last column shows the tree size of the tree generated with the used parameter combination. Shown are means and standard deviation (between brackets) over 5 replicates.

		ABC estimate		
^{10}log (α)	^{10}log (ω)	^{10}log (α)	^{10}log (ω)	Tree size
0	-1	0.27 (0.14)	0.03 (0.08)	3.2 (1.1)
0	1	-0.06 (0.18)	1.02 (0.15)	72.2 (81.21)
0	2	-0.42 (1.51)	1.29 (2.1)	145.6 (104.12)
-0.25	-1	-0.01 (0.09)	0.09 (0.08)	2.8 (1.1)
-0.25	1	-0.26 (0.09)	1.05 (0.19)	19 (12.69)
-0.25	2	-0.22 (0.09)	2.17 (0.34)	21.6 (18.99)
-0.5	-1	0.07 (0.12)	0.02 (0.09)	3.2 (1.79)
-0.5	1	-0.32 (0.12)	0.97 (0.32)	4.8 (1.92)
-0.5	2	-0.42 (0.09)	2.27 (0.49)	7.2 (3.27)

Parameter estimation for the *Lamprologini* phylogeny

Maximum Likelihood estimates for the constant-rates birth-death model

As a reference we inferred speciation and extinction using the constant-rates birth-death model (Nee *et al.* 1994). Using Maximum Likelihood, we obtained an estimate of 0.520 for the speciation rate and extinction was estimated to be 0.

ABC-SMC using literature values on water level changes

When directly sampling from literature values, after 7 iterations of the ABC-SMC algorithm, 443,426,567 proposed values were required to obtain the 10,000 posterior estimates. The low acceptance rate confirms that we are close to convergence (Figure S1). Both sympatric speciation and extinction converge towards a lognormal distribution, whilst allopatric speciation does not seem to converge towards a particular value at all and the final distribution represents the same spread as the prior, suggesting that allopatric speciation has not deviated from the prior (Figure 6.4A). Converting our distributions from 10log to a normal scale, we obtained the following estimates; 1.42 (±0.37) for extinction, 0.52 (±0.06) for sympatric speciation and 16 (±2.75) for allopatric speciation (Table 6.4). The extinction rate is higher than the sympatric speciation rate, but because an extinction event may only remove an incipient species rather than a full species at low water level, the higher extinction rate does not imply a negative diversification rate. When we take a closer look at the correlation between the number of allopatric speciation events and the number of true extinction events (e.g. the number of extinctions in which the species becomes extinct entirely), we find a strong positive relationship (p < 2e-16, R^2 = 0.96) between allopatric speciation events and extinction, indicating that high levels of extinction are coupled with high levels of allopatric speciation. If we remove all particles with extreme allopatric speciation ($\alpha > 1$) from our dataset, the estimated rates are 0.86 (±0.18) for extinction and 0.57 (± 0.07) for sympatric speciation.

ABC-SMC using average rates of water level change

When using an average rate of 10 water level changes per million years the ABC-SMC algorithm ran for 8 iterations, and in the final iteration 848,951,060 values were proposed to obtain the final 10,000 particles. We find lognormal distributions for extinction and sympatric speciation, and estimates for allopatric speciation are generally very small (Table 6.4).

Increasing the rate of water level change to 16 yields similar results, but with a lower sympatric speciation rate. The ABC-SMC algorithm ran 8 iterations, and to obtain the final 10,000 particles, 214,645,333 candidate particles were proposed in the last iteration. Again we find lognormal distributions for both extinction and sympatric speciation, and low values for allopatric speciation (Table 6.4).

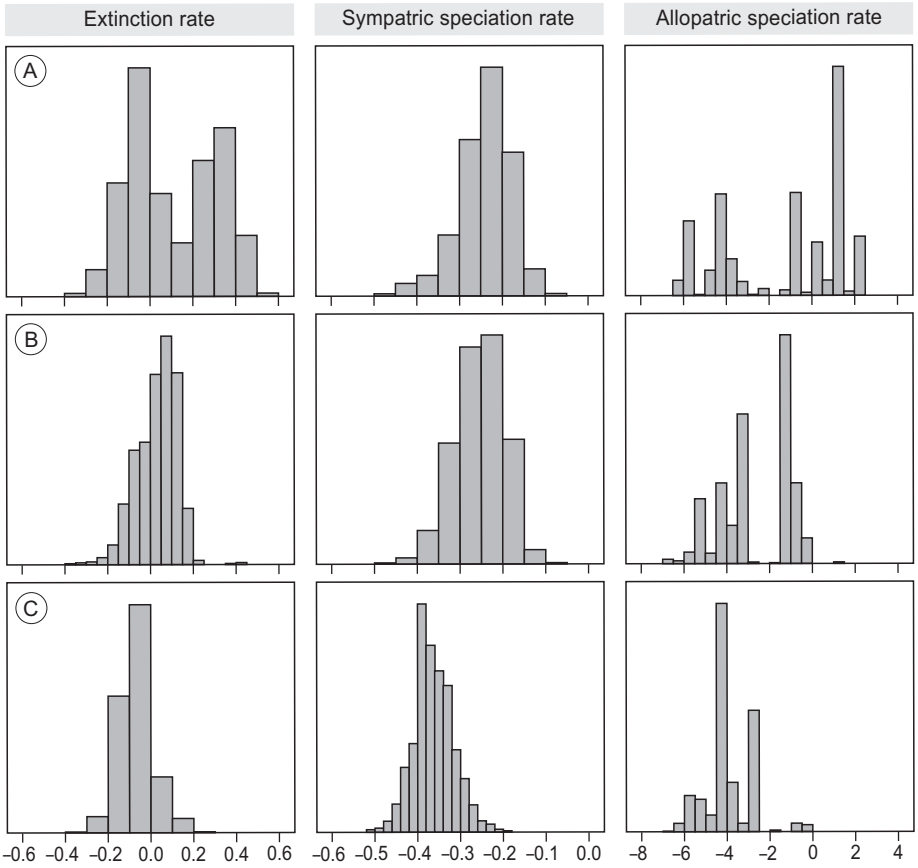

Figure 6.4. Approximate posterior distributions for extinction, sympatric speciation and allopatric speciation rates, fitted to the *Lamprologini* phylogeny, using information from the literature to generate water level changes. A) sampling from literature values B) using a fixed water level rate of 10 C) using a fixed water level rate of 16. x-axis of each histogram is on a ^{10}log scale.

Table 6.4. Mean estimates for the extinction rate(μ), sympatric speciation rate(σ) and the allopatric speciation rate (α) using water level rates (ω) either directly using literature values, or using a mean water level rate based on literature values. Values between brackets represent standard deviations.

ω	M	σ	α
Literature	1.42 (0.37)	0.58 (0.06)	16.37 (2.75)
10	1.10 (0.22)	0.56 (0.07)	0.07 (0.44)
16	0.87 (0.18)	0.44 (0.05)	0.01 (0.06)

ABC-SMC estimating the rate of water level change

In this approach we did not impose a fixed distribution on the water level rate, but rather included the water level rate as a parameter to be inferred during the ABC-SMC process. We obtained our posterior distribution after 9 ABC-SMC steps, and generated 356,988,026 particles to obtain the final distribution of 100,000 particles. Visual inspection of the posterior distributions reveals that they differ from the prior distributions (priors were U(-7,3) (^{10}log scale) for all parameters), and that the posterior distribution of the water level change parameter exhibits a strong bimodal distribution (Figure 6.5A). In all subsequent analyses we analyzed these two peaks (and the associated parameter combinations) separately, where we cut the two peaks in water level change at 10.

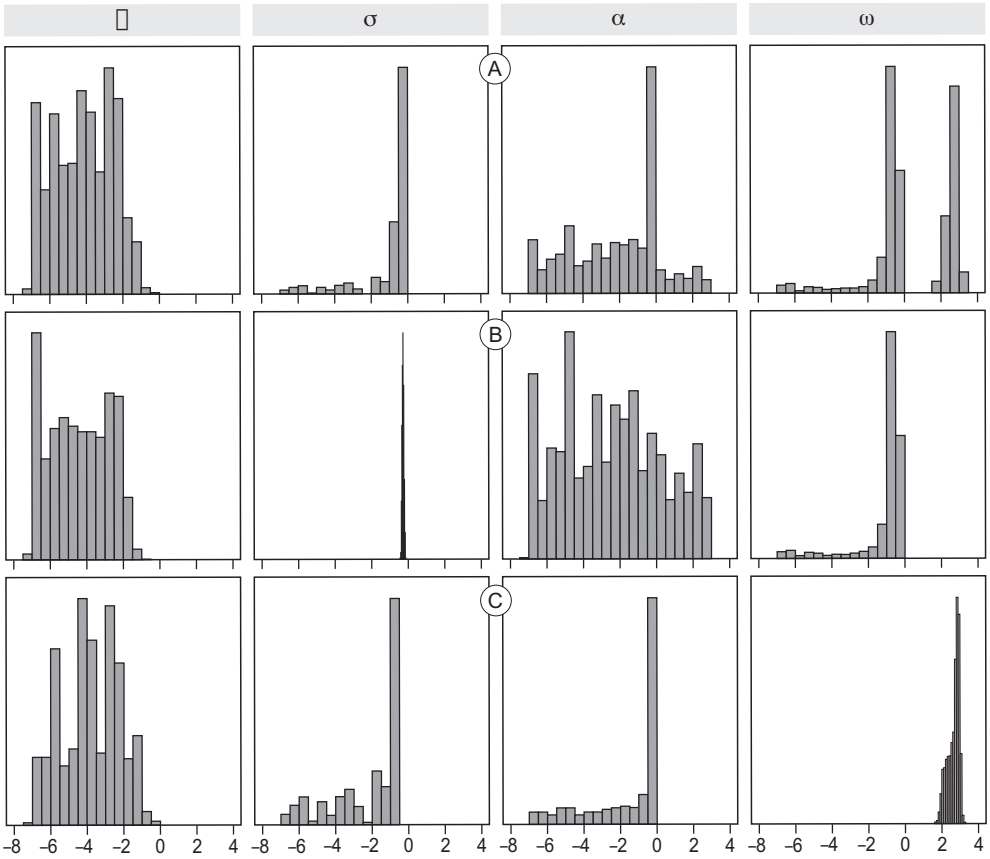

Figure 6.5. Posterior distributions for the extinction rate (μ), the sympatric speciation rate (σ), the allopatric speciation rate (α) and the water level change rate (ω) after 8 iterations of the ABC-SMC algorithm using 100,000 particles per iteration. A) full posterior distribution B) posterior distribution of particles associated with ω) < 10 C) posterior distribution of particles associated with ω > 10. All parameters are plotted on a ^{10}log scale.

The subset of particles associated with water level change < 10 showed a mean rate of water level change of 0.24, which is associated with a very high allopatric speciation rate (28.274). The mean value is misleading here, as Figure 6.5B reveals that the distribution of the Allopatric Speciation parameter resembles a uniform distribution on U(-7,3). Given the low rate of water level change, none of the parameter values for allopatric speciation had any impact on the posterior distribution and no shift from the prior to the posterior distribution was observed. Similarly, the distribution of the extinction rate shows a nearly uniform distribution on the range (-7,-1), which seems to suggest that extinction is close to 0. The sympatric speciation rate distribution resembles a normal distribution, with mean of 0.507 and a small standard deviation (0.039). Parameter estimates are highly similar to the parameter estimates obtained for the constant-rates birth-death model, which resonates the fact that without water level changes, the model reduces to the constant-rates birth-death model.

The subset of particles associated with water level change >10 provides a non-trivial solution (Figure 6.5C), where we do have water level changes, and where the distribution of the allopatric speciation parameter does not resemble the prior. Estimates for the extinction rate cannot be distinguished from 0. The sympatric speciation rate correlated significantly with the allopatric speciation rate (linear regression, slope = -2.48, F = 1.275e06, df = 40956, p < 2.2e-16, R^2 = 0.9689). Thus, with increasing allopatric speciation, sympatric speciation decreased and the two processes were fully complementary. With low sympatric speciation (Table 6.5, $\omega > 10$ & $\sigma \sim 0$), allopatric speciation was comparable to the sympatric speciation rate in the no-water level

Table 6.5. Parameter estimates for the extinction rate (μ), sympatric speciation rate(σ), allopatric speciation rate (α) and the water level rate (ω) after 8 iterations of the ABC-SMC algorithm.. The full distribution resembles the posterior distribution as in figure 6.5A. $\omega < 10$ shows parameter estimates from particles from the posterior distribution in figure 6.5B. $\omega > 10$ shows parameter estimates from the posterior distribution in figure 6.5C. $\omega > 10$ & $\sigma \sim 0$ shows parameter estimates from the full posterior distribution after taking the subset where $\omega > 10$, and the sympatric speciation rate (σ) was smaller than 0.01 (almost 0). $\omega > 10$ & $\alpha \sim 0$ shows parameter estimates from the full posterior distribution after taking the subset where $\omega > 10$, and the allopatric speciation rate (α) was smaller than 0.01 (almost 0). Finally, the constant-rates birth-death estimates show parameter estimates obtained using Maximum Likelihood Estimation on the constant-rates birth-death model (Nee *et al.* 1994). Values between brackets represent the standard deviation.

	μ (per species per MY)	σ (per species per MY)	α (per species per MY)	ω (per MY)
Full distribution	0.004 (0.02)	0.341 (0.211)	16.816 (80.410)	227.45 (335.35)
ω< 10	0.002 (0.01)	0.507 (0.039)	28.274 (103.105)	0.24 (0.17)
ω > 10	0.010 (0.03)	0.103 (0.100)	0.298 (0.253)	554.98 (304.76)
ω > 10 & σ ~ 0	0.010 (0.04)	0.0004 (0.0008)	0.556 (0.043)	783.92 (192.192)
ω > 10 & α ~ 0	0.006 (0.02)	0.220 (0.019)	0.001 (0.002)	290.62 (237.27)
Constant-rates birth-death	0	0.520	NA	NA

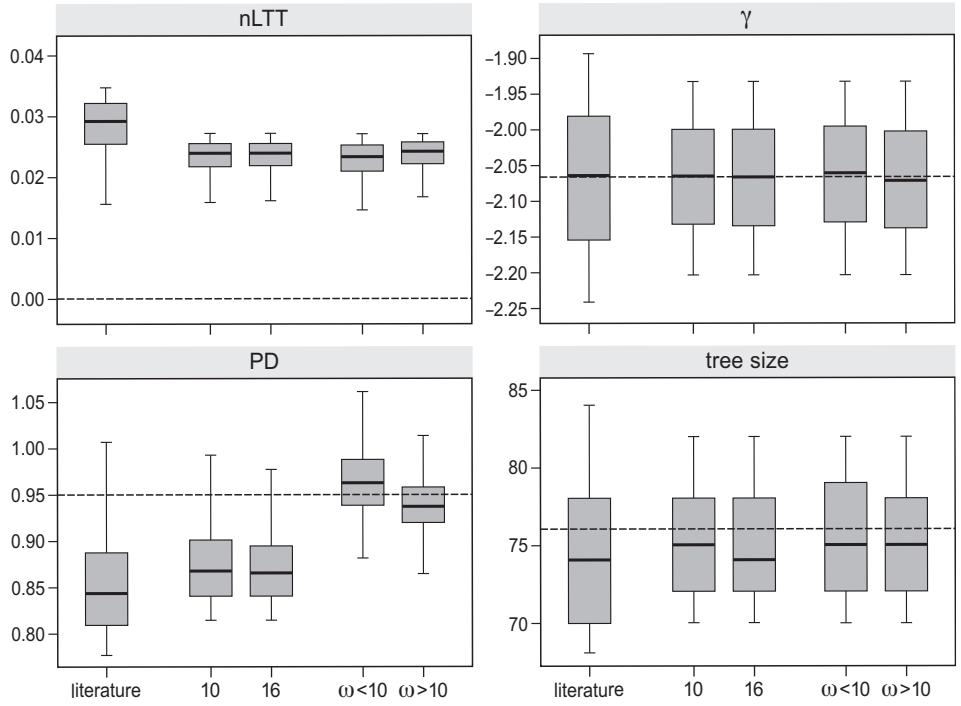

Figure 6.6. Distribution of summary statistics associated with accepted particles in the last iteration for the three scenarios with predefined water rates (either using literature values, an average rate of 10 water level changes per million years or an average rate of 16 water level changes per million years), and for the two found optima when including water level rate as a parameter to be inferred. Dashed line indicates the value of the summary statistic for the *Lamprologini* phylogeny.

change situation ($\alpha = 0.556$). Conversely, with low allopatric speciation (Table 6.5, $\omega > 10$ & $\alpha \sim 0$), estimates for sympatric speciation reveal an intermediate rate of sympatric speciation, of 0.220.

Mean rate of water level change was 555 times per million years, which results in an average waiting time until a water level change event of 1800 years, and an average waiting time between water level drops of 3600 years. If sympatric speciation is low (Table 6.5, $\omega > 10$ & $\sigma \sim 0$), the mean water level rate increases to 784 times per million years (e.g. a water level change event every 1275 years), and with low allopatric speciation rates (Table 6.5, $\omega > 10$ & $\alpha \sim 0$), mean water level rate was 290 times per million years (e.g. a water level change every 3448 years).

Distribution of summary statistics

Comparing the final distributions of summary statistics, especially the distributions of the gamma statistic and the tree size statistic are close to the empirical values (Figure 6.6). Final distributions of the nLTT statistic are comparable and do not overlap with 0. The distribution of the nLTT statistic for the fixed empirical water levels is slightly

higher, but that model converged in 7 iterations compared to the 8 iterations of the other models, and hence the final acceptance threshold for nLTT values was slightly higher. We find most variation among distributions in the phylogenetic diversity statistic, where the distributions of the two optima when also inferring the water level rate most closely approximate the empirical value.

Model validation

For all versions of our model we simulated 1,000,000 phylogenetic trees, and compared summary statistics for these trees with the empirical data. The distributions give an indication of fit of the specific model. Note that this is different than directly looking at the distribution of summary statistics for the particles accepted in the last iteration, as this time we do not select trees to be within an acceptance threshold. The distributions thus show the spread of phylogenetic trees expected when generating with our obtained estimates, and indicate how likely it is to obtain a tree similar to the empirical tree, when using parameter combinations as inferred during ABC-SMC. nLTT values tend to be a bit lower when including the water level parameter in inference, and similarly, phylogenetic diversity estimates tend be closer to the empirical value when including the water level parameter (Figure 6.7). The γ statistic is consistently higher, suggesting that the low γ value of the empirical data is hard to obtain using our model (Figure 6.7). Furthermore, tree size appears to be much higher in the empirical data than in the model. Trees generated by the model often remain without any speciation event, which is especially pronounced when using an average water level rate of 10, in which case the tree size is heavily biased towards low values (Figure 6.7).

Discussion

We have presented a model that infers past speciation and extinction rates, and their interactions with changes in the environment, from a given phylogeny. Previous studies inferring speciation from phylogenetic information have not focused on the interaction with environmental factors. Rather, the focus has been on the dynamics of speciation and the ability to infer extinction rates (Moen & Morlon 2014; Morlon 2014). Furthermore, studies regarding the interplay between allopatric speciation and changes in the environment have focused on a more conceptual analysis of the problem, and did not confront their findings with empirical data (Aguilée *et al.* 2011, 2013). Here we present a model that is able to infer past changes in the environment, and the interaction between these changes and speciation and extinction rates, from phylogenetic data. Using simulated data we were able to detect water level changes from phylogenetic data, and to accurately infer associated allopatric speciation rates. We then applied our model to the cichlid fish tribe of *Lamprologini* to see to what extent past water level changes have shaped current cichlid diversity.

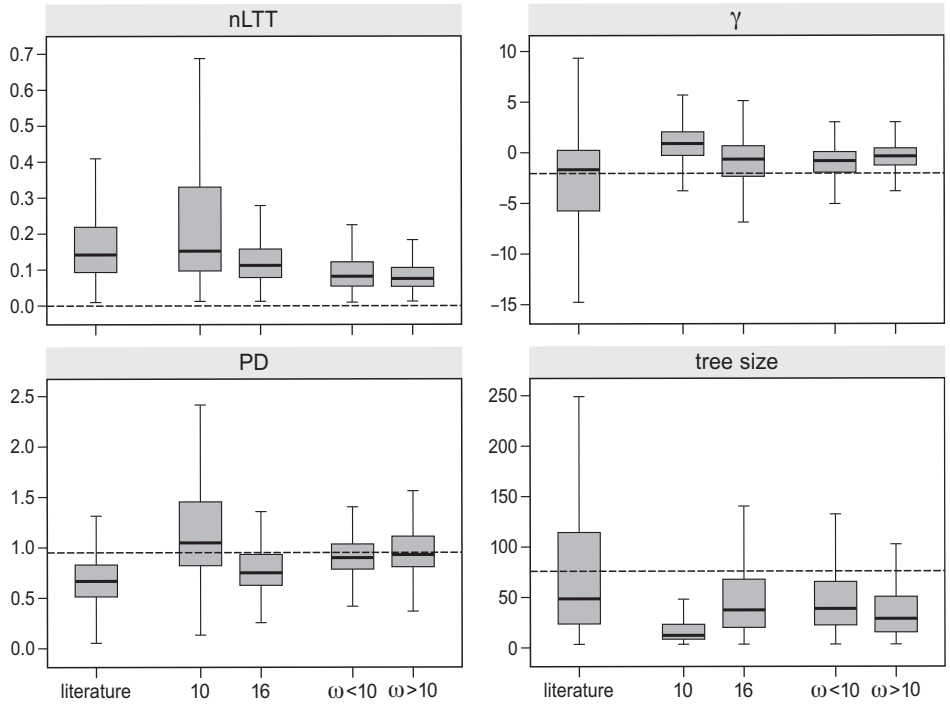

Figure 6.7. Distribution of summary statistics of 1,000,000 trees generated using parameter combinations from the posterior distributions obtained in Figures 6.1 and 6.2. Scenarios correspond to the scenarios in Figure 6.6. Dashed line indicates the value of the summary statistic for the *Lamprologini* phylogeny. Please note that the range of y-axes differs from Figure 6.3.

When using literature values to implement water level changes, we do not recover high levels of allopatric speciation, and estimates for sympatric speciation are close to estimates obtained using the constant-rates birth-death model, which obviously does not incorporate water level changes. Regardless of how we implement the literature values (e.g. either using the mean literature value, or directly using estimated waiting times), allopatric speciation is consistently estimated to be low. Surprisingly we also recover relatively high levels of extinction, although the majority of these extinction events do not lead to true extinction, but rather lead to local extinction at low water level, where the subdivision into two lakes ensures survival of the species in the other lake. When we use our model to also infer the rate of water level change, we recover two competing optima. One optimum without any water level effects and one optimum including water level effects. The optimum without any effects of water level effects again closely matches values inferred using the standard birth-death model (Nee *et al.* 1994). The alternative optimum includes water level effects and infers a zero extinction rate, a high level of water level change, and a negative relationship between sympatric and allopatric speciation. The high rate of water level change ensures that by chance, some of the water level change events line up with branching events in the tree. In

other words: the inferred high water level rates do not necessarily represent reality, but are more likely a result of overfitting the model. Hence, by including water level change as a parameter to be inferred, we recovered two, trivial, solutions: one where water level change is zero, and we recover the constant-rates birth-death optimum, and one where water level change is extremely high, ensuring that at least some water level changes have effect. It thus seems that the phylogeny does not contain enough information in order to infer four different parameters.

The pure birth optimum recovered by our model could also be the result of remnant patterns in the data caused by the tree prior used to construct the phylogeny of *Lamprologini*. The phylogeny of the *Lamprologini* was constructed using BEAST (Drummond *et al.* 2012), using a pure birth prior (Sturmbauer *et al.* 2010). Here, when we include the rate of water level change as a parameter to be inferred, one of the optima we recover is a pure-birth optimum. This suggests that the tree prior used in tree reconstruction can be of potential influence to parameter estimates obtained when using a different model to infer diversification rates. The choice of pure-birth prior might well bias any efforts to infer the extinction rate (even when using matching models) and caution should therefore be taken when inferring diversification rates from a phylogeny using a different model than the model used as tree prior in reconstruction. Elsewhere (Höhna *et al.* 2014b) we found that when using a branch-rate model with estimated and/or low variance, the tree prior used in construction of the phylogeny does not bias any estimates subsequently obtained in diversification analysis. The *Lamprologini* phylogeny was estimated using an autocorrelated lognormal branch-rate prior, which was not tested by Höhna *et al.* 2014. Although we thus can not exclude an effect of the tree prior on our estimates, we do not expect such an effect considering that the autocorrelated lognormal branch-rate prior has low variance.

If water level changes drive allopatric speciation events, the onset of allopatric divergence is at the exact moment that the water level drops. With a water level drop, populations become isolated and start to diverge, if divergence becomes complete and causes speciation, the branching event is identical to the water level change event. It follows that if allopatric speciation is important enough, we expect branching events in a phylogeny to synchronize with water level events, and expect some branching events to occur at exactly the same time. Models used in reconstructing the phylogenetic tree from molecular data explicitly exclude the possibility of simultaneous branching events, and resulting trees therefore cannot exhibit patterns resulting from these simultaneous divergence events. Here we used a model to detect patterns of water level change in a phylogeny, while this phylogeny was reconstructed using a model prior that makes such patterns unlikely. We therefore argue for the inclusion of diversification models incorporating habitat dynamics in tree reconstruction software. Although this need not introduce any significant differences in the tree topology, the distribution of branching times could be substantially influenced, and any subsequent inference focusing on such patterns could be very different. We do realize that including such

models in tree reconstruction software may require incorporation of ABC methods, and even then will be extremely computationally demanding, but we believe that our results justify such an endeavor.

Although we refer in our model to the different implementations of speciation as sympatric and allopatric, care should be taken in interpreting these forms of speciation. We consider here allopatric speciation only on a very large scale, where populations become allopatric over stretches of hundreds of kilometers (Sturmbauer *et al.* 2001). Cichlids are known to be limited in their gene flow over very short distances, where a sand stretch of 50 meters can be enough to bring gene flow between populations to a halt (Rico & Turner 2002). These micro-allopatric speciation events are not captured by the allopatric speciation rate in our model. Rather, these local scale events are captured in our model by sympatric speciation. Hence, sympatric speciation includes not only those speciation events driven by forces that act when two diverging populations occur in full sympatry, but also when two populations are separated by an environmental barrier substantially smaller than the size of the lake, for instance when populations end up on opposite shores (Sturmbauer *et al.* 2001). Allopatric speciation in our model thus solely refers to speciation events caused by geographical isolation over a large distance, driven by changes in water level. All other speciation events are covered by the sympatric speciation rate. Here we recover a low rate of these macro-allopatric speciation events. If this is not the effect of the used tree prior in estimation the phylogeny, the low rate of these macro-allopatric speciation events could either be because these large water level changes have had no effect on speciation rates, or alternatively because they might be overshadowed by micro-allopatric events. As relatively small-scale barriers can already restrict gene-flow and most cichlid species are bound to lacustrine habitats, intermediate water level changes of perhaps only 10 meters could already cause a change in connectivity between populations. Such water level changes would be hard to detect from the phylogeny, as they would not have the same effect on all populations, with some populations experiencing secondary contact whilst other populations start divergence due to introduced separation. Whereas the large scale water level change events cause synchronized divergence, these smaller water level changes only cause divergence in a selective subset of species, if they affect them at all.

Conclusion

Our novel model integrates standard constant-rate birth-death mechanics with environmental changes and speciation induced by geographical isolation. We have analyzed the phylogeny of the tribe of *Lamprologini* in order to see if past water level changes in Lake Tanganyika have contributed to the current diversity of cichlid fish in Lake Tanganyika. Surprisingly our model does not find evidence of past water level changes affecting diversification. The lack of detection of water level changes in the phylogeny of *Lamprologini* could be due to a number of reasons: (1) these large scale water level changes have not had a large impact on the diversification of *Lamprologini*, (2) large

scale water level changes are potentially overshadowed by micro-allopatric and sympatric speciation events, or (3) the tree reconstruction methods used to estimate the phylogeny biased our findings, as these methods do not take into account the possibility of water level events and their effects. Which of these scenarios is most likely remains hard to assess and we suggest to first explore the last scenario by the incorporation of more complex diversification models in tree reconstruction software, in order to remove any biases introduced during tree reconstruction. More detailed genetic studies could further assess the impact of small-scale geological barriers and test to whichat extent changes in the lacustrine habitat due to changes in water level could impact the connectivity between populations. Combined with a thorough assessment of above- and below-water habitat types, this could shed light on how the environment can impact diversity and how changes in the environment can interact with speciation and xtinction in order to shape one of the most diverse vertebrate radiations on this planet.

Supplementary information

Full plots of ABC-SMC progression

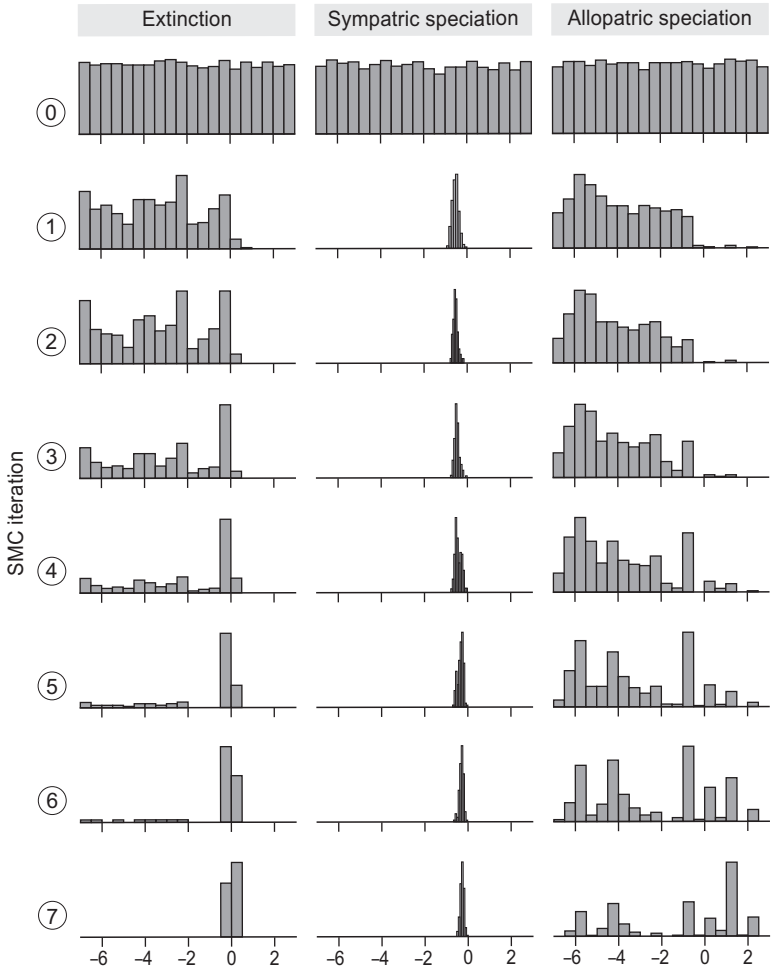

Figure S1. Progression of the ABC-SMC algorithm for the three parameters of interest. Parameters are on a ^{10}log scale (e.g. -2 = 10^{-2} = 0.01), the SMC algorithm starts by sampling from the prior at iteration 0, and from there, using importance re-sampling proceeds to sample a new distribution of 10,000 particles per iteration. With every consecutive iteration, the acceptance thresholds used in the ABC-SMC algorithm are updated (see methods for more details). Rate of water level change consists of the exact values found in the literature.

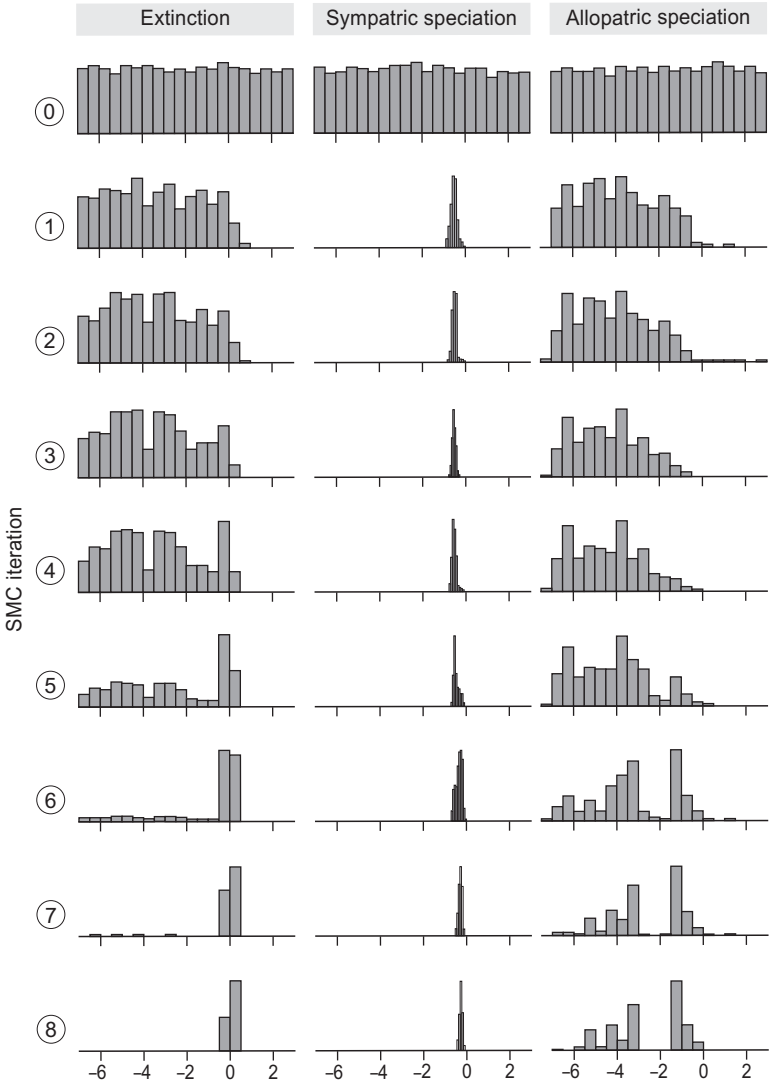

Figure S2. Progression of the ABC-SMC algorithm for the three parameters of interest. Parameters are on a ^{10}log scale (e.g. -2 = 10^{-2} = 0.01), the SMC algorithm starts by sampling from the prior at iteration 0, and from there, using importance re-sampling proceeds to sample a new distribution of 10,000 particles per iteration. With every consecutive iteration, the acceptance thresholds used in the ABC-SMC algorithm are updated (see methods for more details). Rate of water level change is 10 times per million years on average.

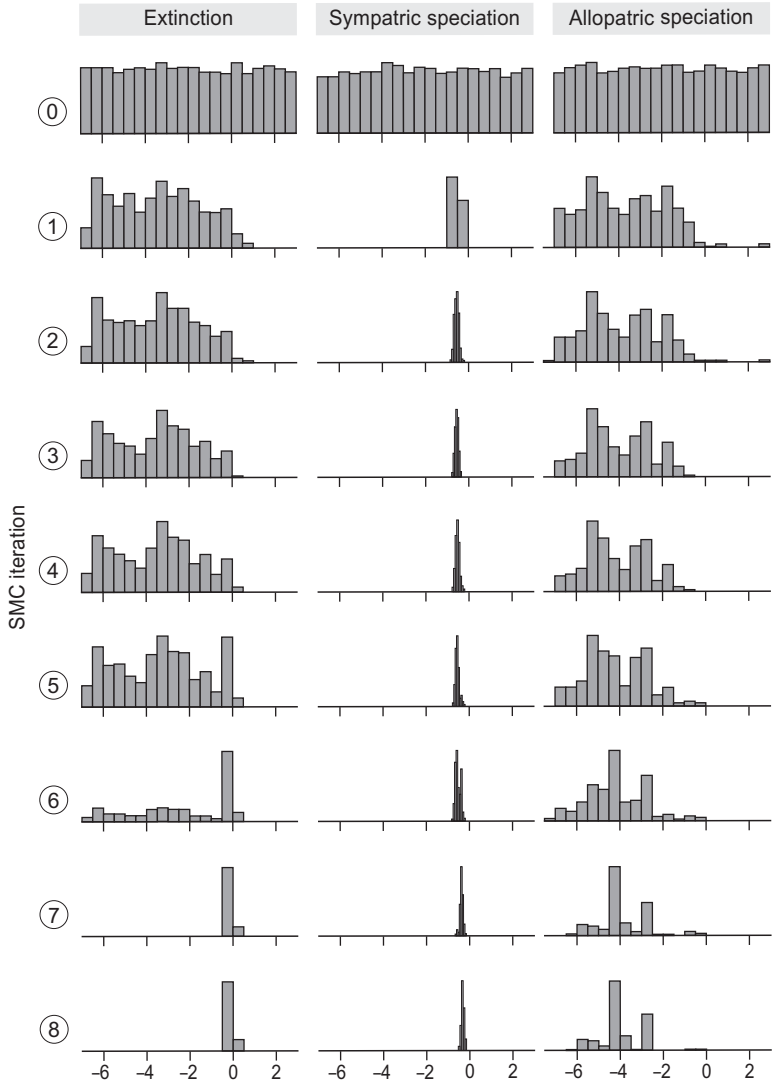

Figure S3. Progression of the ABC-SMC algorithm for the three parameters of interest. Parameters are on a ^{10}log scale (e.g. -2 = 10^{-2} = 0.01), the SMC algorithm starts by sampling from the prior at iteration 0, and from there, using importance re-sampling proceeds to sample a new distribution of 10,000 particles per iteration. With every consecutive iteration, the acceptance thresholds used in the ABC-SMC algorithm are updated (see methods for more details). Rate of water level change is 16 times per million years on average.

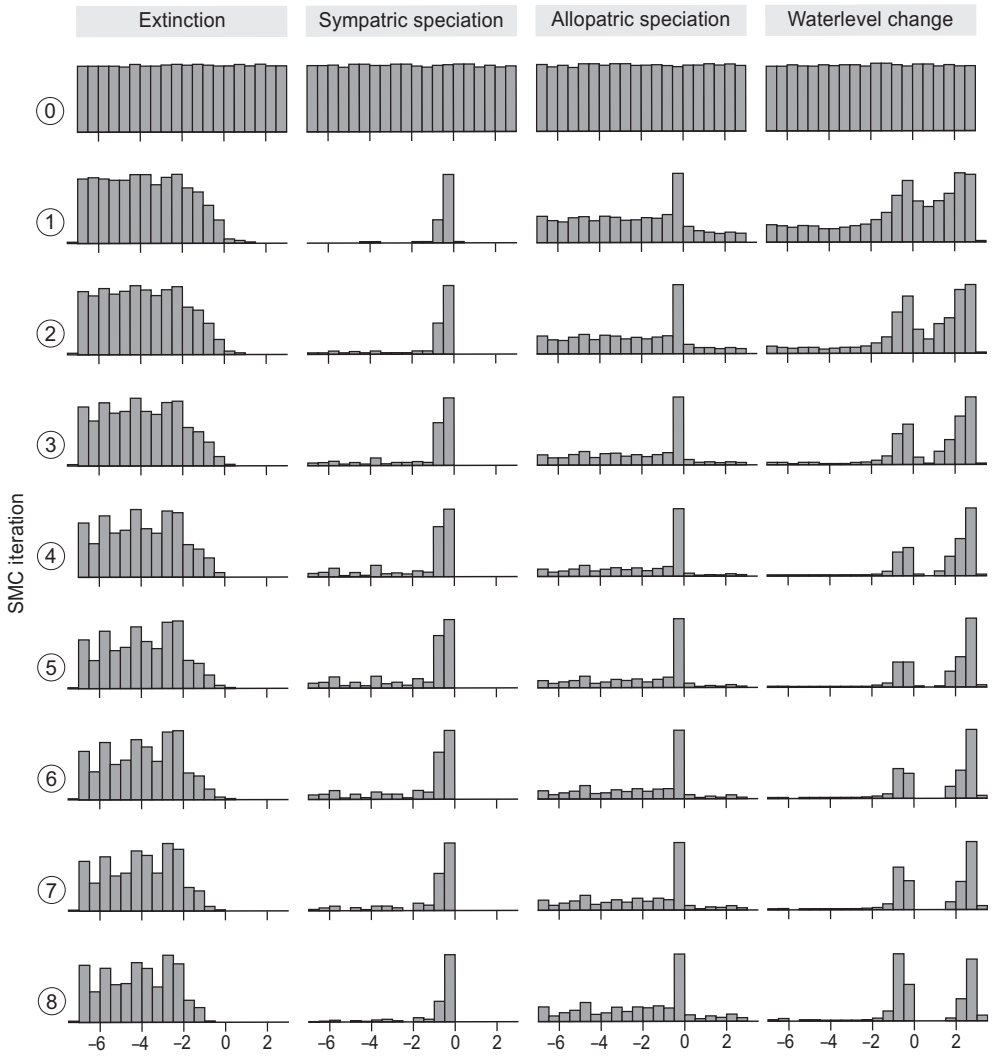

Figure S4. Progression of the ABC-SMC algorithm for the four parameters of interest. Parameters are on a ^{10}log scale (e.g. -2 = 10^{-2} = 0.01), the SMC algorithm starts by sampling from the prior at iteration 0, and from there, using importance re-sampling proceeds to sample a new distribution of 100,000 particles per iteration. With every consecutive iteration, the acceptance thresholds used in the ABC-SMC algorithm are update (see methods for more details).